GANimation: Anatomically-aware Facial Animation from a Single Image

Anmol Mann
University of Victoria
Victoria, BC, Canada
anmolmann@uvic.ca

Abstract

The State-of-the-Art approaches till now could generate expressions out of a discrete set of emotion categories like angry, scary, happy and surprised, etc. but it would be way more interesting if not only we would generate these images but also if we could interpolate between them. This is exactly what GANimation model can do, that is, generating continuous expressions. This means generating an animation between the input real expression and the desired expression.

1. Introduction

The GAN model, [6], implemented in this project can generate anatomically-aware facial animations from a single image under changing backgrounds and illumination conditions. The model is capable of changing expression of a person who has not been seen previously during the training. And this is done just from a single image. To achieve this we have two main components, the first one is the self-learned face attention which can focus into specific parts of the face and the second key component is an anatomically aware expression representation which does not require a 3D face model nor the initialization method, and its trained in an unsupervised manner.

In this work, the overview of the author's approach is discussed. Also, the results generated with this novel approach are observed. Some of the issues in the initial project implementation by the author have also been resolved in this work.

1.1. Motivation

The most interesting fact about this is that all the facial expressions are generated from just one single image and no further information is needed. Furthermore, Action Units are an interesting coding systems and the recent advances in Deep Learning are making great use of it. Therefore, one can expect to see more technologies using them in the near future.

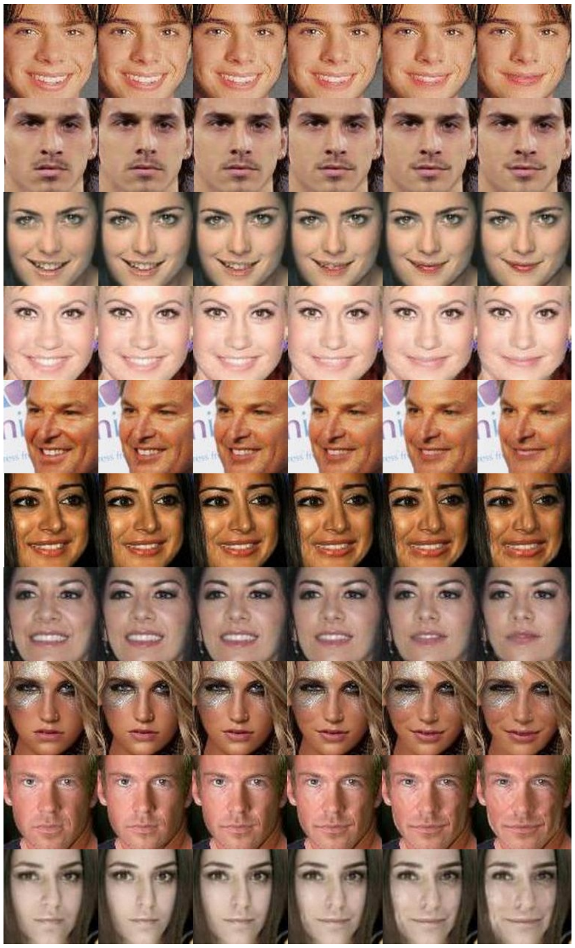

Figure 1. Results obtained are amazing. The images in the leftmost column are the real images provided to the model during testing. Rest all are fake images generated by the Generator.

1.2. Project Goal

Implementing a GAN model that can generate facial animations in the wild and which can be trained in a completely unsupervised manner. Moreover, the model can generate visually compelling smooth images having consistent

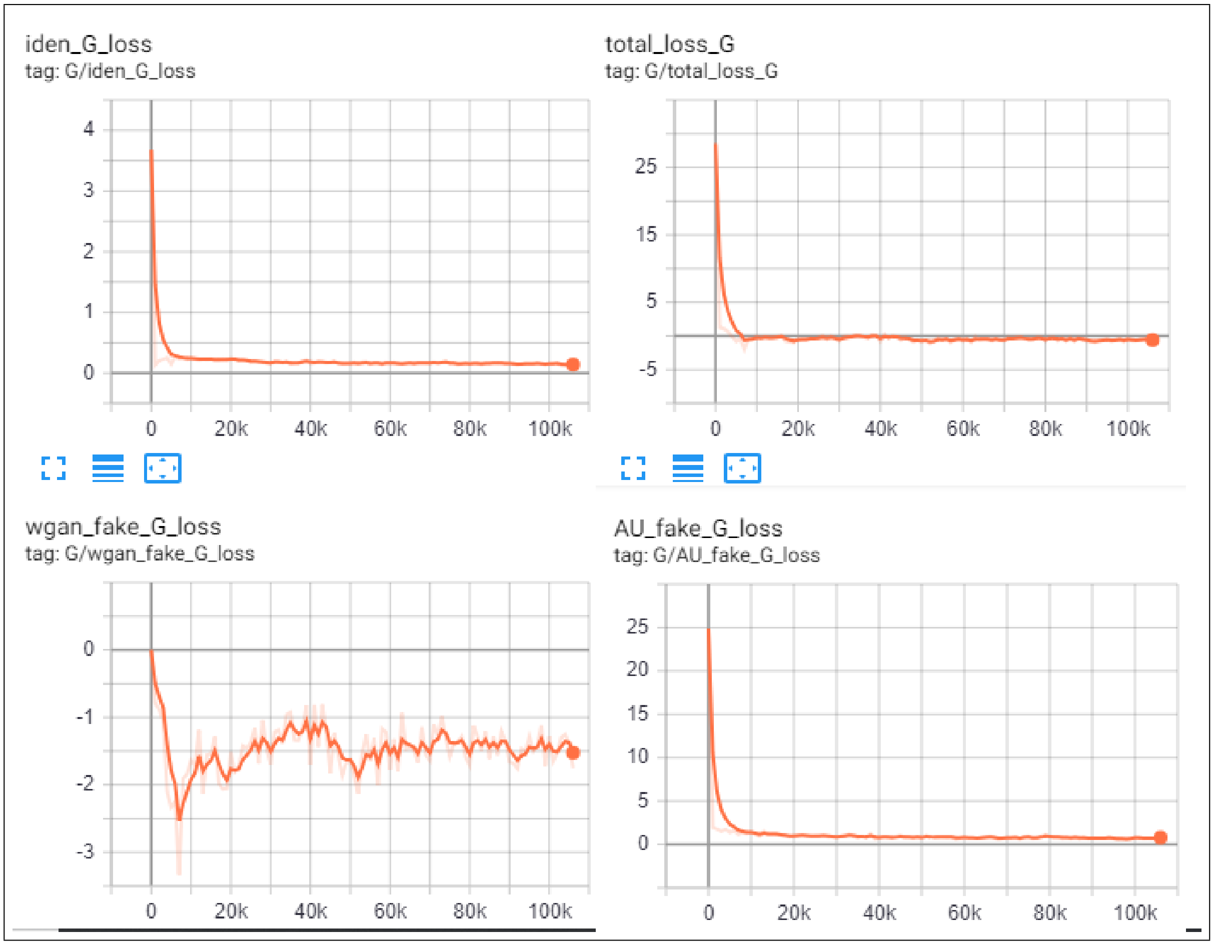

Figure 2. This figure depicts the Generator Loss Values. The graph at the top-right corner is the linear combination of all loss functions for the model's Generator. The graph at top-left corner shows the Generator's identity losses. Bottom-left graph shows the regular WGAN loss [1]. And the bottom-right graph corresponds to the condition-expression loss.

transformations across animations even in some challenging illumination conditions and with low-resolution images as well.

2. Contributions of the original paper

Given an input image of a person under some expression and a desired expression, the novel GAN architecture proposed in [6] can change the original expression into the desired expression while maintaining the person identity all along. As this method has two main components, first one is the expression representation and the second one is the generation mechanism.

This model is not conditioned on discrete emotion categories like State-of-the-Art works but instead it uses Action Units (defined by Facial Action Coding System, FACS) *et al.* [4]. Each Action Unit is directly related to some muscles on the face. So, every expression is encoded as a vector in which every element represents the level of contraction of a muscle in the face. This simply means that the model is conditioned on a one-dimensional AU vector. And, the elements of this vector indicates the presence or absence and the magnitude of that particular Action Unit. The network is trained in an unsupervised manner and only require images with their activated AUs. This is the first key component of this project.

The second key component is the generation mechanism. Normally, all the generators regress the pixel values of the output image but in this model most of the pixels in the output image are directly copied from the input image. And only those pixels which are directly related to the desired expression are changed. To obtain this behavior an attention mechanism is embedded in the model. This attention mechanism is trained in a completely unsupervised manner. Now, every time when we start generating an image, we first detect what pixels need to be changed and how important are these pixels to obtain the desired expression. Basically, the blacker a pixel is, the more important it is to generate

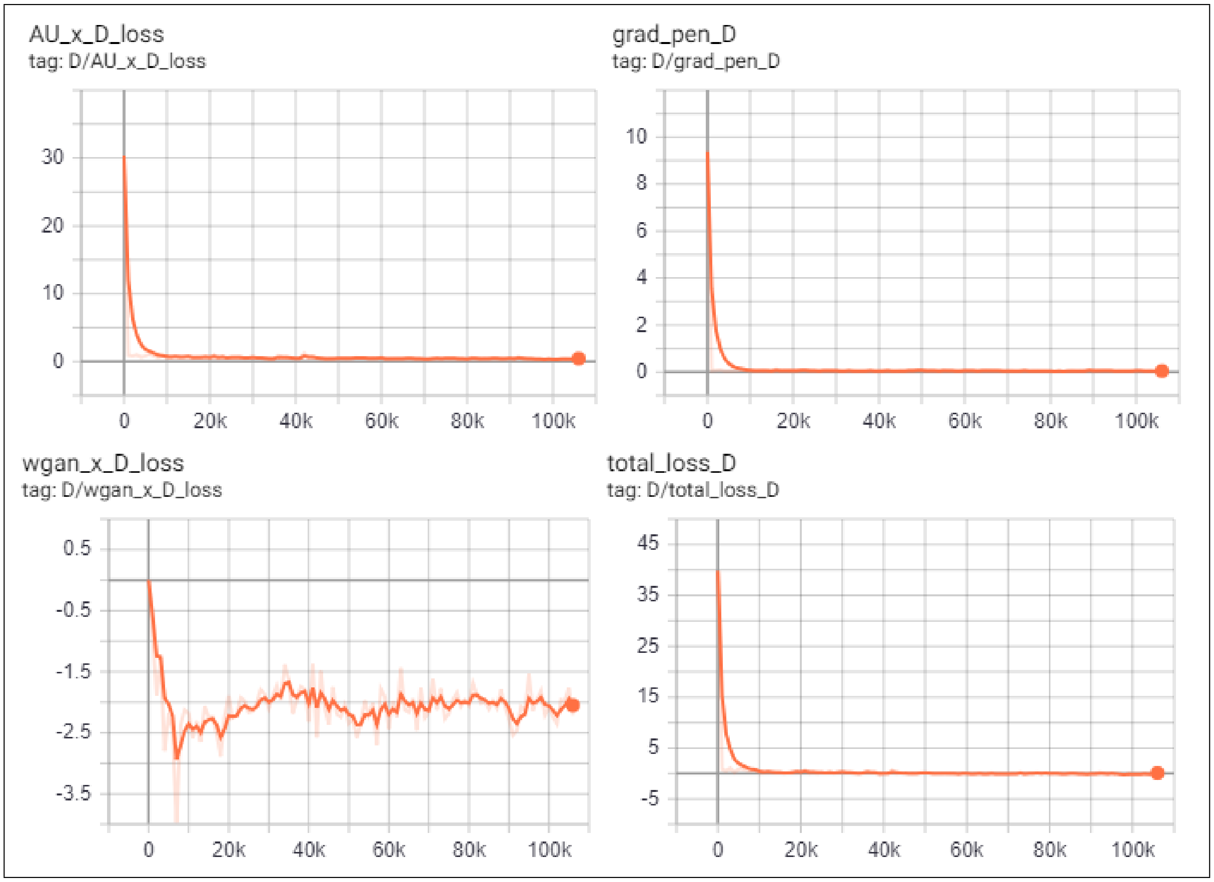

Figure 3. This figure depicts the Discriminator Loss Values. The graph at the bottom-right corner is the linear combination of all loss functions for the model's Discriminator. Bottom-left graph shows the regular WGAN loss [1]. The top-right graph corresponds to the gradient penalty added to improve training of the WGAN as proposed in [5]. And the top-left graph corresponds to the condition-expression loss.

the desired facial expression. This means that the attention mechanism is forcing the generator to only change the pixels with darker activation values in the attention mask. This means that the attention mechanism is forcing the generator to only change the pixels with darker activation values in the attention mask.

During training, the model becomes a bit more complex due to no ground rules present on how the output image should look like. For this several modules are added just for training. These modules are training losses. In GANs, at the beginning everything is just pure noise during the first few iterations. That is when discriminator comes into picture as it forces the model to produce real images. But then minimizing the error becomes insignificant as the attention mask can just simply copy the entire input image to the output. To prevent this, firstly the weight of the attention mechanism is penalized. Secondly, the expression regressor is added to penalize the difference between the desired expression and the generated one. Thirdly, to maintain the identity of the person also along with obtaining the desired expression, a cycle of consistency loss is added which forces the model to recover the original image from the generated one. The loss function obtained at the end is the linear combination of all these partial losses.

3. Contributions of this project

This work is an alternative implementation of the original GANimation project proposed in [6]. One advantage of this project project is that the testing module is more robust and flexible than the initial implementation. The interpolated images can be saved as gif images as well. Also, the interpolated images are generated in a similar manner as demonstrated in [6] and we can also control the number of interpolations we want to generate.

The model is trained on both EmotioNet [3] as well as celebA data-set, rather than just EmotioNet data-set, using the Adam optimizer. Thus, the Action Unit vectors for celebA data-set are first extracted using Openface[1] tool.

[1] Openface2.0: Facial Behavior Analysis Tool [2]

And the results depict that the low resolution images in the celebA data-set are also providing good results during testing.

The normalization layer is a bit different when using Instance Normalization as affine flag needed to be turned off as it was scaling and shifting the normalized data again. This in turn was resulting in no change in the generated expressions and we were getting similar interpolated images with no change in their expressions. Therefore, this issue with instance normalization during model evaluation was resolved in the project's re-implementation.

In the results obtained from this model as in Figure 1, we are not only generating the desired expressions but we are also able to interpolate between them in a realistic way. This is done by controlling the hyper-parameter alpha, which controls the intensity of the Action Unit (AU) vector. These results are cherry-picked as there were also instances where the model failed as well and did not generate desired expressions, as shown in Figure 4. Also, the results generated show how well the model works in different light conditions and with low-resolution images. The loss values obtained by Generator and Discriminator are shown in the Figures 2 and 3.

4. Discussion and future directions

The GAN model implemented in this project generates continuous facial expressions of a person not previously seen during training and only from a single image in an unsupervised manner. The key components of the model are self-learned attention mechanism and anatomically-aware expression representation, using Action Units. The attention layer to handle images under some challenging light conditions so that only those regions of the image are affected that are relevant in the generation of the novel expression. And to dodge the need for multiple training images of a single person with a bunch of different expressions, a two-way generator is used. This makes it possible to transform an image into a desired expression and also, render back the original image from the generated fake image. Therefore, this way we can generate a wider range of expressions and interpolate between them at the same time as well.

There were several instances where the model failed as well as shown in Figure 4. For instance, the model fails when it is provided some extreme input expression or a non-human image. And these failure cases presumably occurred due to the error in the attention mechanism and insufficient data for training purpose. In other words, the model fails to correctly detect which pixels of the image are actually contributing in the generation of the desired expression.

If we could automatically animate the images, it opens doors to many applications in various fields as it we do not need to spend hours with Photo-shop. This technique can help the advancement in the technology used in several areas such as movie industry, fashion and e-commerce business, photography technologies, and many more. As the current model has been tested with the only images as its input. So, one possible future research area for this technique can be in generating video sequences. This needs a slight change in the model architecture as the camera angle comes into picture in this case.

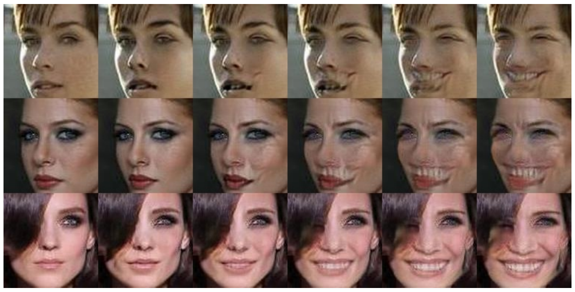

Figure 4. Instances where the model failed. Model does generate extreme uncanny results but it certainly is not able to provide smooth animations.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *CoRR*, abs/1701.07875(4), 2017.

[2] T. Baltrusaitis, A. B. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. *13th IEEE International Conference on Automatic Face Gesture Recognition*, 14(3):59–66, 2018.

[3] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martínez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (6):5562–5570, 2016.

[4] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, 111 15(1):E1454–62, 2014.

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. (5), 2017.

[6] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. *Computer vision - ECCV ... : ... European Conference on Computer Vision : proceedings. European Conference on Computer Vision*, 11214(2):835–851, 2018.